Week 16b: Activity Key

Last Time

- Areal Data Visualization
- Assessing Spatial Structure in Areal Data
- Overview of Areal Data Models

This Time

- Model fitting with Areal Data
- Simulating the spatially correlated areal data
- Modeling continuous spatially correlated areal data

Recall: Disease Mapping

Areal data with counts is often associated with disease mapping, where there are two quantities for each areal unit: Y_i = observed number of cases of disease in county i and E_i = expected number of cases of disease in county i. Generically, this can be modeled as $Y_i|\psi_i \sim Poisson(E_i\psi_i)$.

We will use `S.glm`, `S.CARbym`, and `S.CARleroux` from the `CARBayes` package to fit and compare models using deviance information criteria.

```
formula <- observed ~ offset(log(expected))
no_spatial <- S.glm(formula=formula, data=respiratory_admissions,
                          family="poisson", burnin=10000, n.sample=30000,thin=2, verbose =

print(no_spatial)
```

```
#################
#### Model fitted
#################
Likelihood model - Poisson (log link function)
Random effects model - None
Regression equation - observed ~ offset(log(expected))


#################
#### MCMC details
#################
Total number of post burnin and thinned MCMC samples generated - 10000
Number of MCMC chains used - 1
Length of the burnin period used for each chain - 10000
Amount of thinning used - 2

############
#### Results
############
Posterior quantities and DIC

       Mean       2.5%      97.5% n.effective Geweke.diag
    -0.1644    -0.1829    -0.1457   3978.3000      0.1000

DIC =  2288.246       p.d =  0.9983456       LMPL =  -1149.07
```

```
exp(-.1643)
```

```
[1] 0.8484874
```

```
mean(respiratory_admissions$SMR)
```

```
[1] 0.8605064
```

One way to incorporate spatial structure is with the Besag-York-Mollie (BYM) model, written as

$$
\begin{aligned}
Y_i|\psi_i &\sim Poisson(E_i\psi_i) \\
log(\psi_i) &= x_i^T\beta + \theta_i + \phi_i
\end{aligned}
$$

where we place a CAR prior on ϕ and standard random effects on θ.

$$\begin{aligned}
\phi_k|\phi_{-k}, W, \tau &\sim N(\frac{\sum_{i=1}^k w_{ki}\phi_i}{\sum_{i=1}^k w_{ki}}, \frac{\tau^2}{\sum_{i=1}^k w_{ki}}) \\
\theta_k &\sim N(0, \sigma^2)
\end{aligned}$$

```r
bym <- S.CARbym(formula=formula, data=respiratory_admissions,
                             family="poisson", W=W_mat, burnin=10000,
                n.sample=30000, thin=2, verbose = F)
print(bym)
```

```
#################
#### Model fitted
#################
Likelihood model - Poisson (log link function)
Random effects model - BYM CAR
Regression equation - observed ~ offset(log(expected))


#################
#### MCMC details
#################
Total number of post burnin and thinned MCMC samples generated - 10000
Number of MCMC chains used - 1
Length of the burnin period used for each chain - 10000
Amount of thinning used - 2

############
#### Results
############
Posterior quantities and DIC

               Mean    2.5%   97.5% n.effective Geweke.diag
(Intercept) -0.2201 -0.2458 -0.1931      1700.9        -0.6
tau2         0.3689  0.1803  0.5366        73.5         0.3
sigma2       0.0156  0.0020  0.0577        37.7        -0.2

DIC =  1073.133       p.d =  116.0398       LMPL =  -579.35
```

Alternatively we can specify the following model known as the Leroux model which uses the IAR framework with the ρ term where

$$
\begin{aligned}
Y_i|\psi_i &\sim Poisson(E_i\psi_i) \\
log(\psi_i) &= x_i^T\beta + \phi_i \\
\phi_k|\phi_{-k}, W, \tau &\sim N(\frac{\rho\sum_{i=1}^{k} w_{ki}\phi_i}{\rho\sum_{i=1}^{k} w_{ki} + 1 - \rho}, \frac{\tau^2}{\rho\sum_{i=1}^{k} w_{ki} + 1 - \rho})
\end{aligned}
$$

```
leroux <- S.CARleroux(formula=formula, data=respiratory_admissions,
                             family="poisson", W=W_mat, burnin=10000,
                n.sample=30000, thin=2, verbose = F)
print(leroux)
```

```
#################
#### Model fitted
#################
Likelihood model - Poisson (log link function)
Random effects model - Leroux CAR
Regression equation - observed ~ offset(log(expected))


#################
#### MCMC details
#################
Total number of post burnin and thinned MCMC samples generated - 10000
Number of MCMC chains used - 1
Length of the burnin period used for each chain - 10000
Amount of thinning used - 2

############
#### Results
############
Posterior quantities and DIC

               Mean    2.5%   97.5% n.effective Geweke.diag
(Intercept) -0.2203 -0.2457 -0.1954      3274.3         0.1
tau2         0.3346  0.2210  0.4809      2356.1        -1.7
rho          0.6192  0.3165  0.9147      1824.4        -1.5

DIC =  1074.098       p.d =  117.0745       LMPL =  -584.33
```

Note that the above models result in a single smooth, spatial random surface (defined by the neighborhood structure). The differences in the BYM and the Leroux approaches are fairly minimal.

However, models can also be formulated to incorporate local spatial structure.

One option is the Lee and Mitchell approach, which models the w_{kj} terms rather than setting all to be zero or one. Specifically, an additional variable (Z) is constructed to model dissimilarity between neighboring units. In this case, our z values correspond to the percentage of people defined to be income deprived. Using this value we construct a distance (or dissimilarity) metric between areal units.

Fit this model using `S.CARdissimilarity` and compare to the previous models.

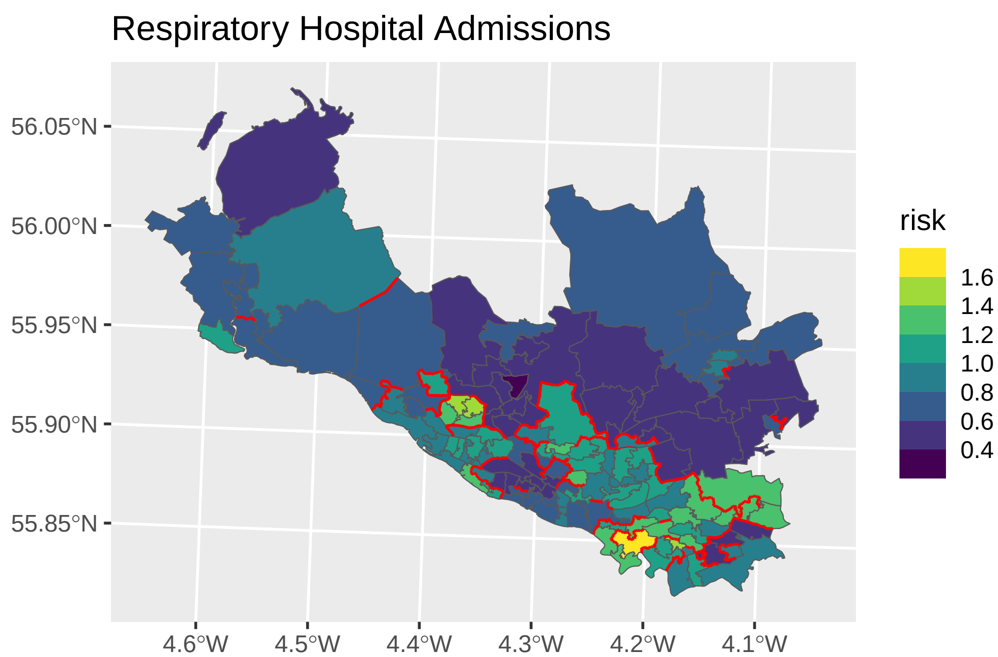
```
income <- respiratory_admissions$incomedep
Z.incomedep <- as.matrix(dist(income, diag=TRUE, upper=TRUE))

dis <- S.CARdissimilarity(formula=formula, data=respiratory_admissions,
                          family="poisson", W=W_mat, Z=list(Z.incomedep=Z.incomedep), verl
                          W.binary=TRUE, burnin=10000, n.sample=30000, thin=2)

print(dis)
```

```
#################
#### Model fitted
#################
Likelihood model - Poisson (log link function)
Random effects model - Binary dissimilarity CAR
Dissimilarity metrics -  Z.incomedep
Regression equation - observed ~ offset(log(expected))


#################
#### MCMC details
#################
Total number of post burnin and thinned MCMC samples generated - 10000
Number of MCMC chains used - 1
Length of the burnin period used for each chain - 10000
Amount of thinning used - 2

############
#### Results
```

```
############
Posterior quantities and DIC

               Mean    2.5%   97.5% n.effective Geweke.diag alpha.min
(Intercept) -0.2196 -0.2413 -0.1981      4781.8        -1.3        NA
tau2         0.1378  0.0970  0.1903      2693.1         0.9        NA
Z.incomedep  0.0499  0.0465  0.0513      2013.6        -0.5    0.0139

DIC =  1058.03       p.d =  99.07121       LMPL =  -564.81

The number of stepchanges identified in the random effect surface
     no stepchange stepchange
[1,]           261         99
```

We can also extract the boundaries, where a stepchange (no neighbor structure) is identified.

```r
border.locations <- dis$localised.structure$W.posterior
respiratory_admissions$risk <- dis$fitted.values /
  respiratory_admissions$expected
boundary.final <- highlight.borders(border.locations=border.locations,
                                sfdata=respiratory_admissions)
st_crs(boundary.final) <- raster::crs(respiratory_admissions)

respiratory_admissions |>
  ggplot() +
  geom_sf(aes(fill = risk)) +
  geom_sf(data = boundary.final, color = 'red') +
  scale_fill_viridis_b() +
  ggtitle('Respiratory Hospital Admissions')
```



Models for continuous data

Now consider a continuous response on areal data. We will use a dataset called `pricedata` on the same areal locations as our previous analysis.

```
library(CARBayesdata)
data(pricedata)
head(pricedata)
```

```
         IZ   price crime rooms sales driveshop     type
1 S02000260 112.250   390     3    68       1.2     flat
2 S02000261 156.875   116     5    26       2.0     semi
3 S02000262 178.111   196     5    34       1.7     semi
4 S02000263 249.725   146     5    80       1.5 detached
5 S02000264 174.500   288     4    60       0.8     semi
6 S02000265 163.521   342     4    24       2.5     semi
```

```
pricedata <- pricedata |>
  mutate(log_price = log(price))
```

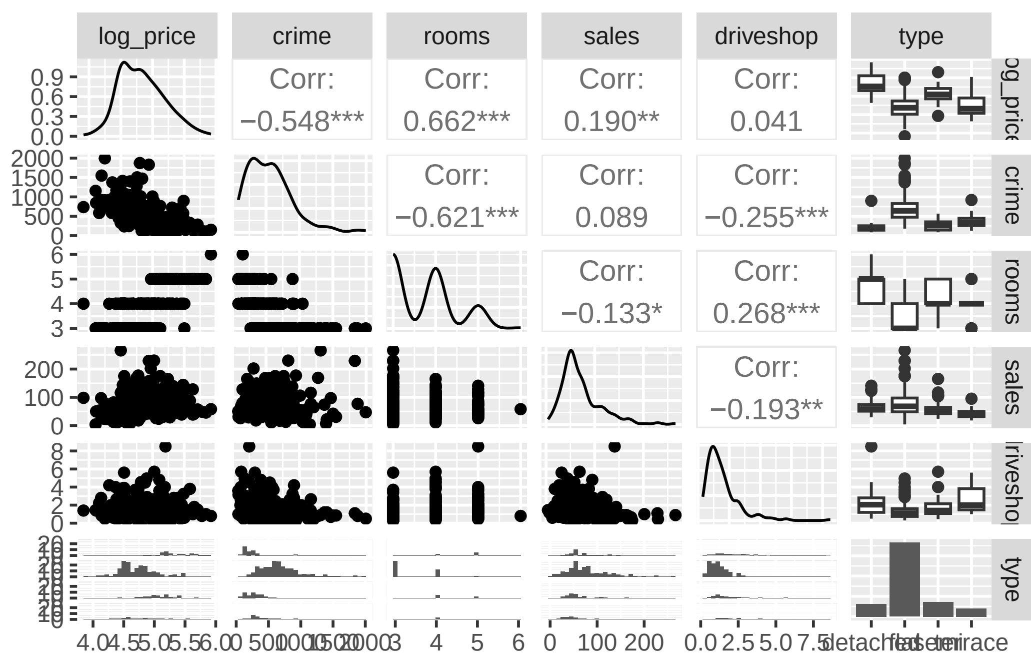
Here is a data dictionary for this dataset:

- **IZ:** The unique identifier for each IZ.
- **price:** Median property price.
- **log_price:** We've created the logarithm of price, which can be useful for modeling given the skewed structure of price.
- **crime:** The crime rate (number of crimes per 10,000 people).
- **rooms:** The median number of rooms in a property.
- **sales:** The percentage of properties that sold in a year.
- **driveshop:** The average time taken to drive to a shopping centre in minutes.
- **type:** The predominant property type with levels: detached, flat, semi, terrace.

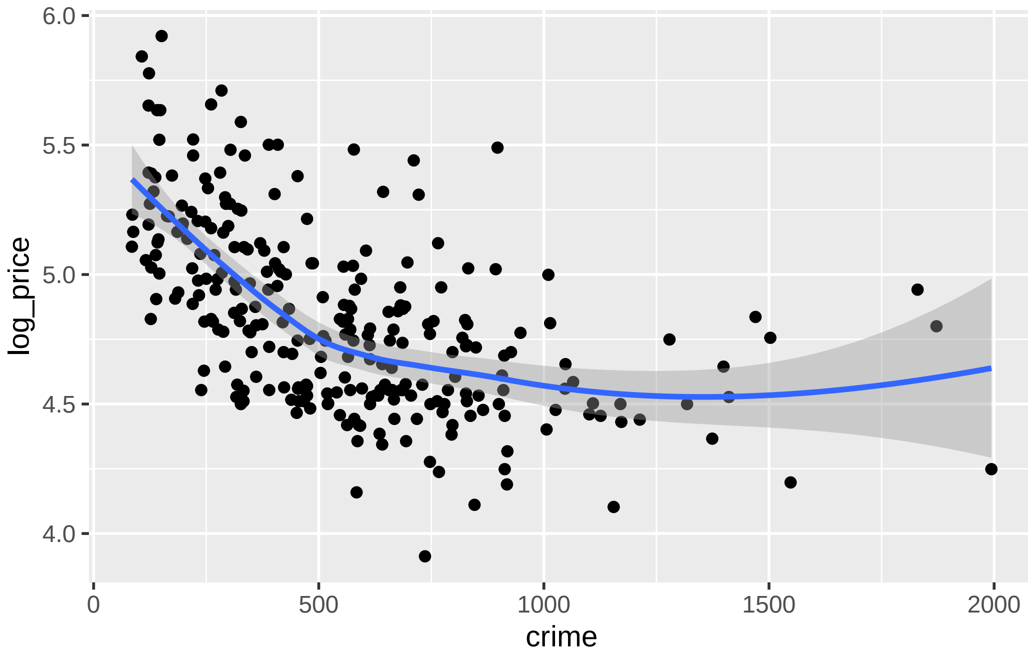
Note that the data curators deleted one observation due to an aberrant value.

Explore mean structure in `log_price` as a function of other variables with data visualization.

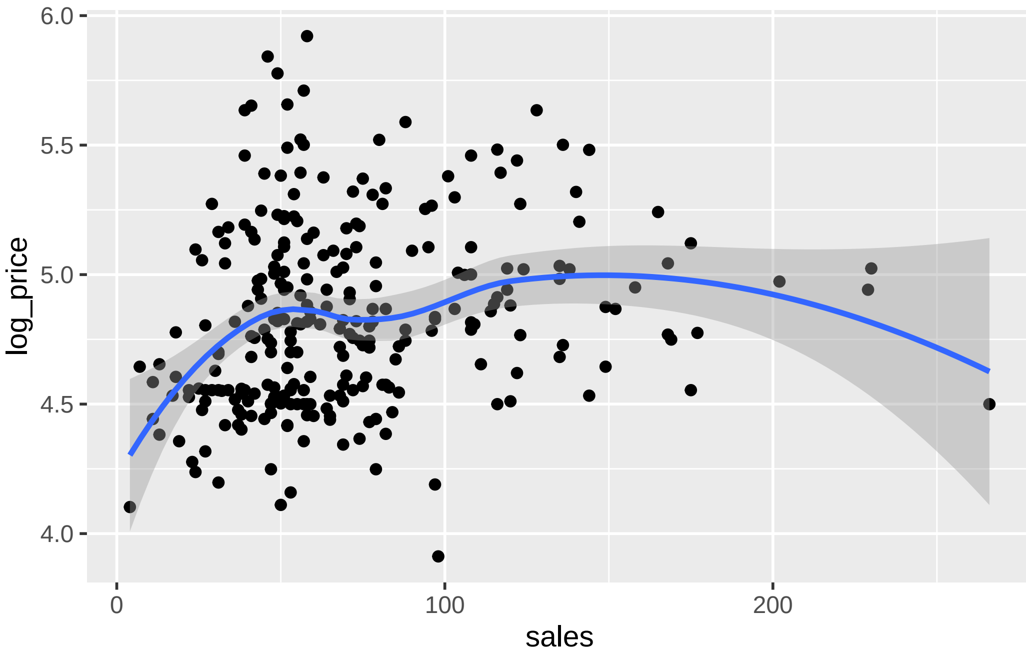
```
library(GGally)
ggpairs(data = pricedata, columns = c(8, 3:7))
```



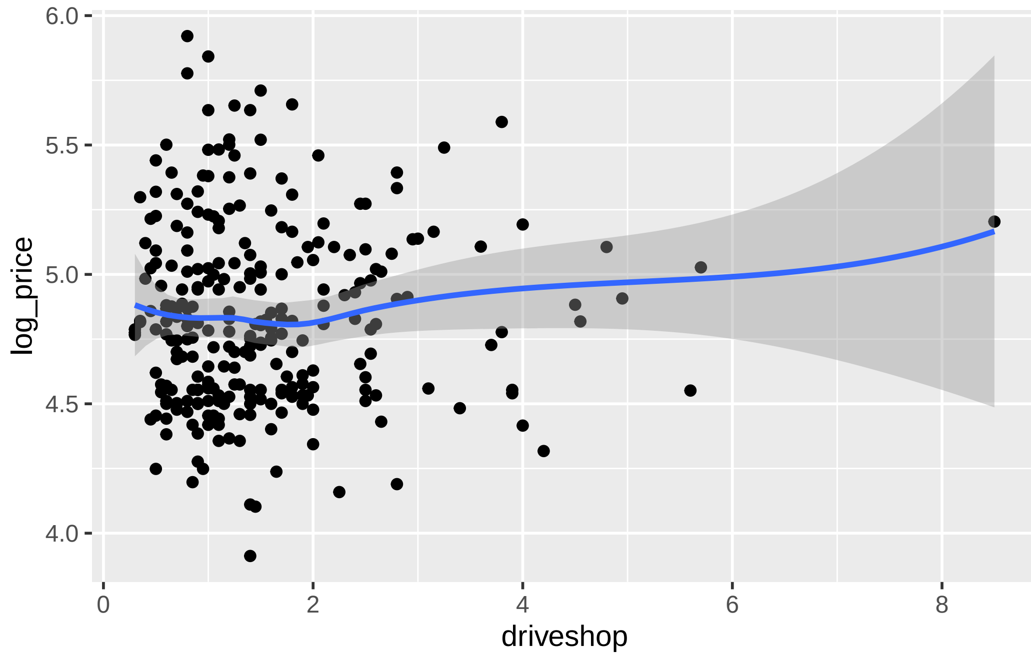
```r
pricedata |>
  ggplot(aes(y = log_price, x = crime)) +
  geom_point() +
  geom_smooth()
```




```r
pricedata |>
  ggplot(aes(y = log_price, x = sales)) +
  geom_point() +
  geom_smooth()
```

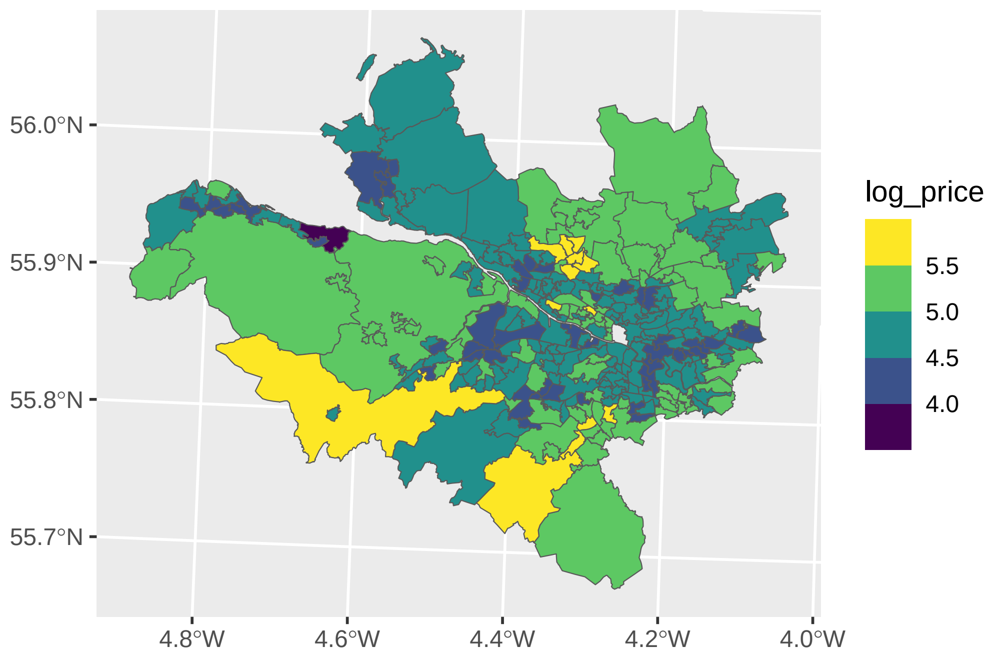



```r
pricedata |>
  ggplot(aes(y = log_price, x = driveshop)) +
  geom_point() +
  geom_smooth()
```



Visualize log price and assess spatial structure

```
comb_price <- GGHB.IZ |>
  right_join(pricedata, by = join_by(IZ))

comb_price |>
  ggplot() + geom_sf(aes(fill = log_price)) +
  scale_fill_viridis_b()
```




```
W_list <- nb2listw(poly2nb(comb_price), style = 'B')
```

```
Warning in poly2nb(comb_price): neighbour object has 2 sub-graphs;
if this sub-graph count seems unexpected, try increasing the snap argument.
```

```
W_mat <- nb2mat(poly2nb(comb_price), style = 'B')
```

```
Warning in poly2nb(comb_price): neighbour object has 2 sub-graphs;
if this sub-graph count seems unexpected, try increasing the snap argument.
```

```
moran.test(comb_price$price, W_list, alternative = 'two.sided')
```

```

    Moran I test under randomisation

data:  comb_price$price
weights: W_list

Moran I statistic standard deviate = 12.289, p-value < 2.2e-16
alternative hypothesis: two.sided
```

```
sample estimates:
Moran I statistic       Expectation          Variance
      0.449887576      -0.003717472       0.001362464
```

```r
geary.test(comb_price$log_price, W_list, alternative = 'two.sided')
```

```
    Geary C test under randomisation

data:  comb_price$log_price
weights: W_list

Geary C statistic standard deviate = 6.143, p-value = 8.097e-10
alternative hypothesis: two.sided
sample estimates:
Geary C statistic       Expectation          Variance
      0.685454074       1.000000000       0.002621829
```

```r
pull_resids <- lm(log_price~poly(crime,2) + rooms + poly(sales,2) + factor(type) + driveshop

moran.mc(residuals(pull_resids), W_list, 1000)
```

```
    Monte-Carlo simulation of Moran I

data:  residuals(pull_resids)
weights: W_list
number of simulations + 1: 1001

statistic = 0.2871, observed rank = 1001, p-value = 0.000999
alternative hypothesis: greater
```

```r
geary.mc(residuals(pull_resids), W_list, 1000)
```

```
    Monte-Carlo simulation of Geary C

data:  residuals(pull_resids)
weights: W_list
number of simulations + 1: 1001
```

```
statistic = 0.75125, observed rank = 1, p-value = 0.000999
alternative hypothesis: greater
```

Implement a statistical model for log price that includes spatial correlation

```r
lm1 <- S.CARleroux(log_price~ crime + rooms + sales + factor(type) + driveshop, data=comb_pri
 burnin=10000, n.sample=30000, thin=10, n.chains=1, verbose = F)
```

```
Warning in mat2listw(W, style = "B"): neighbour object has 2 sub-graphs
```

```r
print(lm1)
```

```
#################
#### Model fitted
#################
Likelihood model - Gaussian (identity link function)
Random effects model - Leroux CAR
Regression equation - log_price ~ crime + rooms + sales + factor(type) + driveshop


#################
#### MCMC details
#################
Total number of post burnin and thinned MCMC samples generated - 2000
Number of MCMC chains used - 1
Length of the burnin period used for each chain - 10000
Amount of thinning used - 10

############
#### Results
############
Posterior quantities and DIC

                       Mean    2.5%   97.5% n.effective Geweke.diag
(Intercept)          4.1386  3.8687  4.3974      2000.0        -0.1
crime               -0.0001 -0.0002 -0.0001      2000.0         0.5
```

```
rooms                  0.2326  0.1864  0.2809      2000.0        -0.3
sales                  0.0023  0.0017  0.0030      2000.0         1.8
factor(type)flat      -0.2948 -0.4016 -0.1874      1838.6        -0.8
factor(type)semi      -0.1719 -0.2707 -0.0720      2000.0         0.6
factor(type)terrace -0.3229 -0.4383 -0.2026      2137.4        -0.4
driveshop              0.0033 -0.0296  0.0366      1496.9         0.3
nu2                    0.0227  0.0119  0.0328       441.1        -1.7
tau2                   0.0524  0.0237  0.0932       384.9         1.8
rho                    0.9139  0.7463  0.9922       738.4        -0.4

DIC =  -159.8958       p.d =  102.9693       LMPL =  58.19
```

```r
lm2 <- S.CARleroux(log_price~poly(crime,2) + rooms + poly(sales,2) + factor(type) + drivesho
 burnin=10000, n.sample=30000, thin=10, n.chains=1, verbose = F)
```

```
Warning in mat2listw(W, style = "B"): neighbour object has 2 sub-graphs
```

```r
print(lm2)
```

```
#################
#### Model fitted
#################
Likelihood model - Gaussian (identity link function)
Random effects model - Leroux CAR
Regression equation - log_price ~ poly(crime, 2) + rooms + poly(sales, 2) + factor(type) +
    driveshop


#################
#### MCMC details
#################
Total number of post burnin and thinned MCMC samples generated - 2000
Number of MCMC chains used - 1
Length of the burnin period used for each chain - 10000
Amount of thinning used - 10

############
#### Results
############
Posterior quantities and DIC
```

```
                      Mean    2.5%   97.5% n.effective Geweke.diag
(Intercept)         4.2468  3.9953  4.4955      1651.2        -0.1
poly(crime, 2)1    -0.9087 -1.4738 -0.3217      1767.9        -0.4
poly(crime, 2)2     0.4919  0.0676  0.9650      1697.6         1.0
rooms               0.2195  0.1688  0.2699      2000.0         0.2
poly(sales, 2)1     1.5455  1.1416  1.9451      2000.0        -0.8
poly(sales, 2)2    -0.4201 -0.8067 -0.0594      2000.0         0.3
factor(type)flat   -0.2640 -0.3776 -0.1532      1595.6         1.0
factor(type)semi   -0.1620 -0.2634 -0.0626      1872.3        -1.3
factor(type)terrace -0.2908 -0.4151 -0.1661      1766.3        -0.2
driveshop           0.0021 -0.0329  0.0364      1337.9        -0.9
nu2                 0.0230  0.0128  0.0324       439.7         0.7
tau2                0.0482  0.0217  0.0890       434.4        -0.8
rho                 0.9221  0.7503  0.9921       760.0         0.5

DIC =  -158.6826       p.d =  99.61987       LMPL =  61.72
```

```r
lm3 <- S.CARleroux(log_price~poly(crime,2) + rooms + sales + factor(type) + driveshop, data=
 burnin=10000, n.sample=30000, thin=10, n.chains=1, verbose = F)
```

```
Warning in mat2listw(W, style = "B"): neighbour object has 2 sub-graphs
```

```r
print(lm3)
```

```
#################
#### Model fitted
#################
Likelihood model - Gaussian (identity link function)
Random effects model - Leroux CAR
Regression equation - log_price ~ poly(crime, 2) + rooms + sales + factor(type) + driveshop


#################
#### MCMC details
#################
Total number of post burnin and thinned MCMC samples generated - 2000
Number of MCMC chains used - 1
Length of the burnin period used for each chain - 10000
Amount of thinning used - 10
```

```
############
#### Results
############
Posterior quantities and DIC

                       Mean    2.5%   97.5% n.effective Geweke.diag
(Intercept)          4.0795  3.8233  4.3167      2000.0         0.6
poly(crime, 2)1     -1.0610 -1.6159 -0.4842      1841.8        -0.6
poly(crime, 2)2      0.4649 -0.0124  0.9344      1761.9         0.6
rooms                0.2211  0.1687  0.2735      2000.0        -1.2
sales                0.0023  0.0016  0.0029      1829.2         0.7
factor(type)flat    -0.2571 -0.3661 -0.1454      2000.0         0.2
factor(type)semi    -0.1648 -0.2614 -0.0622      2000.0        -0.1
factor(type)terrace -0.3054 -0.4274 -0.1805      2000.0         0.9
driveshop            0.0016 -0.0321  0.0368      1409.0         0.2
nu2                  0.0242  0.0146  0.0335       517.4         1.8
tau2                 0.0458  0.0192  0.0839       384.4        -1.1
rho                  0.9244  0.7734  0.9915       872.4         0.2

DIC =  -147.3871       p.d =  95.4976       LMPL =  58.85
```

```r
lm4 <- S.CARleroux(log_price~crime + rooms + poly(sales,2) + factor(type) + driveshop, data=
 burnin=10000, n.sample=30000, thin=10, n.chains=1, verbose = F)
```

```
Warning in mat2listw(W, style = "B"): neighbour object has 2 sub-graphs
```

```r
print(lm4)
```

```
#################
#### Model fitted
#################
Likelihood model - Gaussian (identity link function)
Random effects model - Leroux CAR
Regression equation - log_price ~ crime + rooms + poly(sales, 2) + factor(type) + driveshop


#################
#### MCMC details
#################
```

```
Total number of post burnin and thinned MCMC samples generated - 2000
Number of MCMC chains used - 1
Length of the burnin period used for each chain - 10000
Amount of thinning used - 10

############
#### Results
############
Posterior quantities and DIC

                       Mean    2.5%   97.5% n.effective Geweke.diag
(Intercept)          4.2828  4.0300  4.5387      1435.7        -3.0
crime               -0.0001 -0.0002  0.0000      1516.5         0.7
rooms                0.2335  0.1869  0.2814      1984.8         2.7
poly(sales, 2)1      1.5866  1.1638  1.9961      2154.6        -1.0
poly(sales, 2)2     -0.3958 -0.7769 -0.0134      2000.0         1.8
factor(type)flat    -0.3002 -0.4138 -0.1866      2000.0         1.4
factor(type)semi    -0.1674 -0.2668 -0.0645      1613.1         0.7
factor(type)terrace -0.3119 -0.4345 -0.1895      2000.0         1.2
driveshop            0.0047 -0.0294  0.0396      1049.7         1.8
nu2                  0.0213  0.0094  0.0312       375.6         0.2
tau2                 0.0560  0.0256  0.1047       340.1        -0.1
rho                  0.9073  0.7154  0.9905       536.8         0.9

DIC =  -174.7421       p.d =  107.8794       LMPL =  66.23
```